Linguistically Motivated Parallel Parsebanks

Helge Dyvik[♠], Paul Meurer[♡],
Victoria Rosén[♠♡], and Koenraad De Smedt[♠♡]

[♠]University of Bergen, Sydnesplassen 7, N-5007 Bergen (Norway)
[♡]Uni Digital, Allégaten 27, N-5007 Bergen (Norway)

{dyvik|paul.meurer|victoria|desmedt}@uib.no

Abstract

Parallel grammars and parallel treebanks can be a useful method for studying linguistic diversity and commonality. We use this approach to study how arguments to similar predicates are realized across languages. To that end, we formulate formal principles for aligning at phrase and word levels based on translational correspondences at predicate-argument level. A first version of a new tool for creating, storing, visualizing and searching treebank alignment at different levels has been constructed.

1 Introduction

A central concern within theoretical linguistics is the discovery of unifying patterns behind language diversity. Our aim is to study linguistic diversity and commonality through the use of parallel grammars and parallel treebanks. By parallel grammars we mean grammars for different languages constructed according to common principles, as in the ParGram project [3]. When parallel grammar writing is anchored in a common Lexical-Functional Grammar (LFG) framework with theory-imposed constraints [1], differences in grammatical analyses across languages are likely to reflect real differences between languages, rather than accidentally different descriptive strategies among grammarians.

A parallel treebank is a syntactically and possibly semantically analyzed parallel corpus, aligned not only on the word and sentence levels, but also on intermediate levels. Parallel treebanks, in which translationally corresponding phrases are linked, are a valuable (and still rare) resource, for example for research on innovative combinations of memory-based and knowledge-based machine translation.

Parallel treebank construction is a recently established and rapidly developing field, and it already includes experiments in automatic phrase alignment, notably Samuelsson and Volk [11]. Some problems arise from the fact that the syntactic structures in the treebanks to be aligned sometimes reflect different principles of analysis. Whereas Samuelsson and Volk's method starts from *n*-gram alignment

(i.e. from the identification of translational correspondences between strings of words) to support phrase alignment, we want to explore the opposite direction: starting from correspondences between predicate-argument structures in a pair of sentence-aligned and tentatively word-aligned monolingual treebanks constructed from the two sides of a parallel (translational) corpus, we derive alignments at the phrase and word levels. In particular, we want to pursue the following hypothesis: *On the basis of monolingual treebanks constructed from a parallel corpus by means of parallel grammars, it will be possible to achieve automatic word and phrase alignment with significantly higher precision and recall than hitherto achieved through other means.*

In an LFG analysis a given f-structure (functional structure) is typically associated with more than one node in the c-structure (constituent structure). A set of nodes projecting the same f-structure is said to constitute a *functional domain*. Assuming an alignment of subsidiary f-structures, we expect that automatic phrase alignment can be achieved by alignments among the nodes in the functional domains of corresponding f-structures, according to criteria spelled out in 3.2 below.

Further assuming that the c-structures are organized according to common principles, the aligned phrasal categories are expected to be typologically informative. Our aim is to test these assumptions on typologically diverse languages: Norwegian, Dutch, Tigrinya and Georgian. Maximal c-structure diversity is guaranteed by the fact that these languages are spread out on the configurationality continuum — the configurational languages Norwegian and Dutch are at one end, and the free word order language Georgian is at the other [7], with Tigrinya in between [8].

In the current phase of our research, we are testing this approach on test suites especially constructed to bring out differences among the languages in the mapping of arguments to syntactic functions. Below we discuss the principles of our methodology, we propose formal alignment principles, and we present the first version of a tool that is unique in that we create, store, visualize and search correspondences between multiple languages at two syntactic levels, c- and f-structure, through a Web interface.

2 Methodology

An LFG-based parsebanking approach [9] is extended to multiple languages. The LFG framework is used because it is a substantial theory about the class of possible human languages, and not just a tool for grammatical description. While the basic projection formalism for codescription in LFG allows a wide range of c-structures to be associated with the same f-structure, and vice versa, including pairings that are implausible linguistically, recent LFG research has proposed strong universal constraints on the possible relationship between the two kinds of structures, implying empirical claims about the limits of possible variation among languages.

A central contribution to this research is Bresnan's development of a theory of constraints on the relationship between c-structures and f-structures [1]. By basing

our grammars on Bresnan's proposals we intend to extend the notion of parallel grammars from just considering f-structure, as in ParGram, to encompassing c-structure as well. A consequence of adopting such common principles is that we will approach a situation where phrase-structural differences between analyses of translationally corresponding sentences will reflect genuine differences among the languages, and not just arbitrarily different principles of analysis among grammarians.

A fundamental task of a grammatical theory is to account for the way in which form and meaning are linked in natural languages, for example showing how phrases in a sentence pick out the participants and their roles (agent, patient, beneficiary etc.) in described situations. Within LFG, the link between syntactic phrases and participant (or argument) roles is mediated by an inventory of syntactic functions like SUBJ, OBJ, OBJth, OBL, XCOMP, etc.; this can be seen as a formalized version of well-known concepts from traditional grammar. There are cross-linguistic constraints on the way in which the arguments of a given predicate are mapped to syntactic functions in the f-structure, and there is a body of research that attempts to establish what these constraints are. The developing theory, Lexical Mapping Theory (LMT) [2, 6], has had to be continually revised as the number of typologically diverse languages investigated has increased. Tigrinya and Georgian raise different non-trivial problems for the application of LMT and hence provide a challenging and fruitful basis for approaching the questions of universal constraints on argument linking.

A multilingual parallel treebank provides information about the set of syntactic functions which is crosslinguistically available for a given argument. Our hypothesis is that LMT will allow the derivation of (possibly underspecified) information about the semantic role of an argument from such sets of alternative syntactic functions that can realize it. In a sense, semantic roles would then be labeled by their sets of alternative syntactic expressions in a way analogous to how alternative translations express the semantic properties of words in the Semantic Mirrors approach [5]. If the hypothesis is confirmed, this would be a highly interesting result both theoretically and practically. Our parallel treebank will provide a unique opportunity to approach these hypotheses, and similar ones, on a solid empirical basis.

3 Alignment principles

3.1 The intuitive notions

Parallel corpora are traditionally aligned on the sentence and word levels. On the sentence level, the default expectation is that all source sentences are aligned to one or (more rarely) more than one target sentence, and vice versa. Only complete sentence-formed omissions or additions in the translations lead to unaligned sentences.

On the word level, on the other hand, a source word is ideally aligned to a

target word (or a target word sequence) if and only if they correspond translationally, where 'translational correspondence' in a text does not only mean semantic correspondence. Rather, words should be aligned only if their surroundings, too, correspond. 'Surroundings' in this connection must be defined in terms of the syntactically expressed argument structure of the sentence, within which the position of the word kan be determined.

Thus, a source word W_S and a target word W_T are taken to correspond translationally only if (i) W_T can in general (out of context) be taken to be among the semantically plausible translations of W_S, i.e., W_T belongs to the set of 'linguistically predictable translations' (LPT) of W_S, and (ii) W_S and W_T occupy corresponding positions within corresponding argument structures. Henceforth we will refer to criterion (i) as 'LPT-correspondence' between a source and a target word. We assume that source/target nouns also enter into 'LPT-correspondence' with target/source pronominal forms. It remains to make precise the criteria for saying that a source argument structure and a target argument structure 'correspond'; this is discussed in 3.2.

Phrase level correspondence is intermediate between sentence level correspondence and word level correspondence. The intuition behind the notion of translational correspondence on the phrase level is that a source phrase Ph_S and target phrase Ph_T are taken to correspond if (i) they contain corresponding words, (ii) Ph_S contains no word or phrase corresponding to a target word or phrase outside Ph_T, and similarly (iii) Ph_T contains no word or phrase corresponding to a source word or phrase outside Ph_S.

Given the criteria sketched above for word-level correspondence, we need not state as an independent criterion that corresponding phrases must express corresponding (in some sense) argument structures (or argument structure parts); they necessarily will if their words truly correspond. Furthermore, we do not require that Ph_S and Ph_T also occupy corresponding positions within translationally corresponding argument structures, as we assume on the level of word correspondences. Such a requirement might put too strong demands on the degree to which translationally corresponding sentences must also correspond syntactically in order for phrase and word alignment to occur. However, position within corresponding argument structures should be a criterion for ranking competing alternative phrase alignments.

This explication of the criteria for correspondence implies that there is a mutual dependence between correspondences on the different levels of a source/target sentence pair. Therefore we cannot first achieve full word-level alignment and only then go on to consider correspondences on the higher levels, nor can we do the opposite. Rather, we need a non-monotonic bootstrapping approach to alignment: based on a seed of an initial tentative, possibly partial and many-to-many word-level alignment relation, the alignment of argument structures takes place, which in turn may justify further word level alignments and also eliminate some of the initial alignments. We assume that this procedure will reach a fixed point. The linking of c-structure phrase nodes only takes place when such a fixed point has been

reached and the word-level alignment is stable and one-to-one.

3.2 Phrase alignment based on parallel LFG analyses

In an LFG analysis, the argument structure properties of a phrase *Ph* are expressed in the value of PRED in the f-structure F which *Ph* 'projects' by the mapping function ϕ (see [4], ch. 4), in conjunction with other properties of F. The value of PRED is always a 'semantic form'. Let $L(Pr)$ be the lexical expression of the predicate Pr in a semantic form $Pr\langle ARG_1 \ldots ARG_n \rangle$.[1] Furthermore, let $P(ARG_i)$ be the predicate in the semantic form of the f-structure to which ARG_i is argument-linked,[2] let $P(ADJ)$ be the predicate in the semantic form of an adjunct ADJ,[3] and let $F^{-\phi}$ be the set of c-structure nodes projecting the f-structure F.

A source f-structure F_S is said to 'correspond' to a target f-structure F_T if F_S and F_T have partially or fully corresponding PRED-values such that $PRED_S = Pr_S\langle ARG_{1S} \ldots ARG_{nS} \rangle$ and $PRED_T = Pr_T\langle ARG_{1T} \ldots ARG_{mT} \rangle$, where

(i) the number of arguments n and m may or may not differ,

(ii) there is LPT-correspondence between $L(Pr_S)$ and $L(Pr_T)$,

(iii) for each ARG_{iS}, there is LPT-correspondence between $L(P(ARG_{iS}))$ and either some $L(P(ARG_{j_T}))$ or some $L(P(ADJ_T))$ of an ADJ_T in F_T, and, conversely,

(iv) for each ARG_{iT}, there is LPT-correspondence between $L(P(ARG_{iT}))$ and either some $L(P(ARG_{j_S}))$ or some $L(P(ADJ_S))$ of an ADJ_S in F_S,

(v) the LPT-correspondences can be aligned one-to-one, and

(vi) there is no adjunct ADJ in F_S such that $L(P(ADJ))$ is word-aligned with a target node projecting an f-structure outside F_T, and vice versa for adjuncts in F_T.

This includes the special case when F_S and F_T have fully corresponding PRED-values $PRED_S = Pr_S\langle ARG_{1S} \ldots ARG_{nS} \rangle$ and $PRED_T = Pr_T\langle ARG_{1T} \ldots ARG_{nT} \rangle$, where

(i) the PRED-values have the same number of arguments $ARG_1 \ldots ARG_n$,

(ii) there is LPT-correspondence between $L(Pr_S)$ and $L(Pr_T)$,

[1] Example: In the f-structure for the sentence *John sleeps*, the semantic form is '$\mathsf{sleep}\langle(\uparrow \text{SUBJ})\rangle$', and $L(\mathsf{sleep})$ is the word form *sleeps*.

[2] Example: In the semantic form '$\mathsf{sleep}\langle(\uparrow \text{SUBJ})\rangle$' from the previous footnote, ARG_1 is argument-linked to the SUBJ, whose semantic form is 'John', which is hence the value of $P(ARG_1)$. The value of $L(P(ARG_1))$, then, is the word form *John*.

[3] Since the PRED-value of an adjunct may be supplied by a preposition, this definition must be sharpened to pick out the semantic form of the OBJ of the preposition in such cases. Thus, if ADJ is the f-structure of an adjunct *on the table*, $P(ADJ)$ would be the semantic form 'table'.

(iii) for every i, $1 \leq i \leq n$, there is LPT-correspondence between $L(P(ARG_{iS}))$ and $L(P(ARG_{iT}))$,

(iv) the LPT-correspondences can be aligned one-to-one, and

(v) there is no adjunct ADJ in F_S such that $L(P(ADJ))$ is word-aligned with a target node projecting an f-structure outside F_T, and vice versa for adjuncts in F_T.

Unlike this special case, the general case does not ensure meaning equivalence between the corresponding f-structures, since it allows corresponding arguments to occur in different orders in the semantic forms. This leaves the degree of semantic equivalence between translationally corresponding complex expressions to some extent open as an empirical question, and also exempts grammar writers from the requirement of achieving completely uniform cross-linguistic criteria for argument ordering, both of which freedoms we consider desirable.

In cases of null pronominal arguments, $L(P(ARG_i))$ is undefined, since there is no lexical expression of the argument. In such cases we assume that the requirement of LPT-correspondence is satisfied if there is a corresponding argument or adjunct in the other language (according to the other criteria above) which is either also a null pronominal argument or has a lexical expression which is not aligned with anything else.

In cases where a source f-structure corresponds to more than one target f-structure, or vice versa, the alternatives are ranked according to (i) the closeness of the correspondence (the special case above being closer than cases involving adjuncts, for example), and (ii) the occurrence vs. non-occurrence of the corresponding f-structures within corresponding embedding f-structures, where cases of corresponding embedding structures take priority.

Once corresponding f-structures have been identified according to the criteria above, all and only the word-alignment links which are in accordance with the f-structure correspondences are kept. In particular, this limits the alignment of nouns with pronouns (including null pronominals) to those cases which are motivated by the surrounding argument structures.

Now phrase alignment can be defined based on the correspondence relation between source and target f-structures and the concomitant word alignments. The set of nodes given by $F^{-\phi}$ constitutes a functional domain within the c-structure. All nodes within a functional domain are alignment candidates. However, we clearly cannot link all nodes in a source functional domain $F_S^{-\phi}$ to all nodes in the corresponding target functional domain $F_T^{-\phi}$. The reason is that as we move downwards in the functional domain along the head path in the c-structure, we may leave behind sister nodes contributing arguments and adjuncts to the shared f-structure (e.g. a subject NP as we move from the S mother to the VP daughter). But aligned nodes should only dominate corresponding material. Furthermore, in cases of long-distance dependencies there may be such contributing nodes that are not dominated within the functional domain at all. Hence, for a source node n_S, we need to make

sure that we only align it with such target nodes n_T that dominate corresponding material (we do not align a source VP which does not dominate the subject NP with a target S dominating the translation of the source subject, even though the source VP and the target S project corresponding f-structures).

Given two corresponding f-structures F_S and F_T, this can be done by the following procedure. For every node n_S in $F_S^{-\phi}$ and every node n_T in $F_T^{-\phi}$, find the set $LL(n_S)$ of linked lexical nodes dominated by n_S (i.e., lexical nodes which are word-aligned with something in the target string), and find the set $LL(n_T)$ of linked lexical nodes dominated by n_T. Align n_S and n_T if and only if $LL(n_S)$ and $LL(n_T)$ are non-empty, all the nodes in $LL(n_S)$ are aligned with nodes in $LL(n_T)$, and vice versa.

Notice that these definitions leave open the possibility that the source or the target phrase may contain material, such as further adjuncts (but not further arguments), not corresponding to anything in the target or source, respectively. Given the frequency of additions and omissions in translations, we need that latitude.

4 The parallel treebanking tool

4.1 Functionality

To help our scientific exercise, we have built initial extensions of the LFG PARSEBANKER [10] to support parallel parsebanking. To our knowledge, there is no prior tool that adequately allows the creation, storage, visualization and search of translational correspondences at multiple levels of structure in parallel treebanks through a Web interface. We briefly describe its current functionality.

Since we currently do not have the bilingual resources required for the automatic performance of the initial tentative word alignment between our languages, the alignment of f-structures and words is presently done manually. The tool allows us to do this for corresponding sentences by dragging the index of a subsidiary source f-structure onto the index of the corresponding target f-structure. The alignment information is stored in a database as an additional layer. We envision doing this automatically in the future.

The procedure for the subsequent alignment of c-structure nodes presented in 3.2 is implemented in the tool, taking the manually aligned f-structures, with their concomitant word-alignments, as input.

4.2 Examples

We will illustrate by means of a few examples. In the screenshots in Figures 1–3 the Norwegian and Georgian subsidiary f-structures[4] have been aligned manually according to the criteria given in 3.2. The f-structure correspondences are shown in the indices on the substructures: an index of the form $\boxed{n \rightarrow m}$ tags structure

[4]The f-structures are shown in 'PREDs-only mode', i.e., many grammatical features, irrelevant to present purposes, have been suppressed.

n and indicates that it is aligned with structure *m*. The automatically derived c-structure alignments are shown by the curved lines. Nodes which share alignments are connected by heavy lines, and the alignment is marked only on the top member of such a set in order not to clutter up the representation unnecessarily. Dotted lines indicate distinct functional domains.

While Norwegian, like English, expresses the beneficiary either as an oblique prepositional phrase or an NP in a double object construction, Georgian only offers the latter possibility. A simple example of an alignment of the two constructions is provided in 1.

(1) (a) *Georg ga en bok til Katarina.*
George gave a book to Catherine
'George gave a book to Catherine.'

(b) *gia-m eḳa-s ṭign-i misca.*
George-ERG Catherine-DAT book-NOM he-gave-it-to_her
'George gave Catherine a book.'

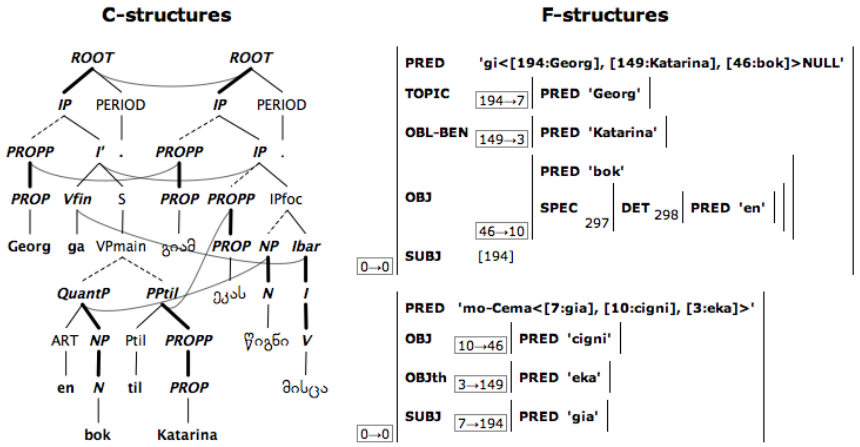

Figure 1: Screenshot of two-level alignment for Example 1

In Figure 1, the Norwegian SUBJ is aligned with the Georgian SUBJ (since the Norwegian SUBJ is unified with the TOPIC, the latter is also aligned with the Georgian SUBJ), the Norwegian OBJ is aligned with the Georgian OBJ, and the Norwegian OBL-BEN is aligned with the Georgian OBJth. As a consequence of the last-mentioned alignment the c-structure nodes PPtil and PROPP are aligned. Furthermore, we may notice that the two grammars happen to have the three arguments of *give* in different orders in the semantic forms, but this does not prevent alignment according to the criteria in 3.2.

Example 2 shows a case where an adjunct in Norwegian corresponds to an argument in Georgian.

(2) (a) *Også på denne konvolutt-en stod navn-et hennes.*
Also on this envelope-DEF stood name-DEF her
'Her name was on this envelope, too.'

(b) *am ḳonverṭ-sa-c mis-i saxel-i eçera.*
this-DAT envelope-DAT-too her-NOM name-NOM it-was-written.
'Her name was on this envelope, too.'

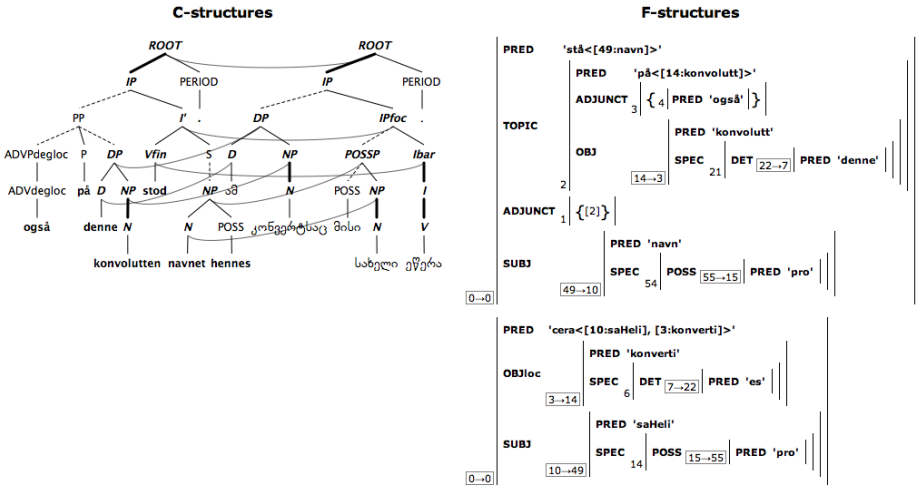

Figure 2: Screenshot of two-level alignment for Example 2

In the Norwegian f-structure in Figure 2, the TOPIC is identical with a member of ADJUNCT, and the OBJ of this shared value is aligned with OBJloc in the Georgian f-structure. As a result, the Norwegian DP under PP and the Georgian DP under IP are aligned.

A Norwegian-Georgian example involving a long-distance dependency (topicalization) in Norwegian, but not in Georgian, is shown in 3.

(3) (a) *Georg antar jeg du mener.*
George assume I you mean.
'George I assume you mean.'

(b) *vpikrob, rom gias gulisxmob.*
I-assume-it, that George-DAT you-mean-him
'I assume you mean George.'

In the Norwegian c-structure in Figure 3, the nodes ROOT, IP, PERIOD, I', Vfin (the upper one), S and VPmain belong to the same functional domain, projecting the entire f-structure. Still, the nodes I', S and VPmain do not enter into any

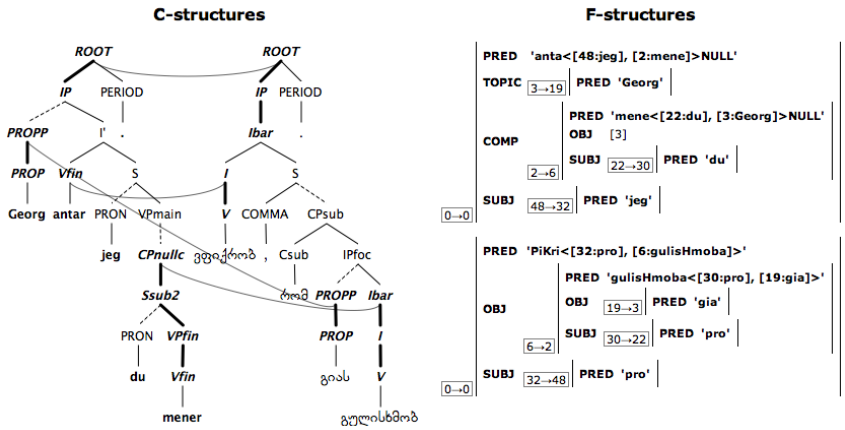

Figure 3: Screenshot of two-level alignment for Example 3

alignment because they do not dominate the word *Georg*, while the latter is word-aligned with a word dominated by the Georgian nodes ROOT, IP, Ibar (the upper one) and S, which would have been the alignment candidates.

On the other hand, the Norwegian nodes CPnullc, Ssub2, VPfin and Vfin (the lower one) are aligned with the Georgian nodes Ibar, I and V (the lower ones). The latter nodes dominate a verb (*gulisxmob*, 'you-mean-him') expressing the second person subject by inflection, while no overt pronoun in the sentence corresponds to the Norwegian second person pronoun *du*, which is therefore unaligned. Hence the criteria are satisfied for aligning both the Norwegian phrase *du mener* and the single verb *mener* with the phrase nodes dominating the single Georgian verb *gulisxmob*; see 3.2 on the treatment of lexically unexpressed PREDs like 'pro'.

However, the same Norwegian nodes CPnullc, Ssub2, VPfin and Vfin are *not* aligned with the Georgian nodes CPsub and IPfoc, although the latter belong to the same functional domain as the nodes Ibar, I and V below them. The reason is that CPsub and IPfoc dominate the name form *gias*, whose Norwegian word-alignment partner *Georg* is not dominated by CPnullc, Ssub2, VPfin or Vfin.

4.3 Search

The LFG Search tool [10] is being extended with a parallel search mode. For aligned sentence pairs, certain c-structure nodes and f-structure nodes (that is, subsidiary f-structures) will be aligned. To make alignment relations searchable, an alignment relation has been introduced as shown in Example 4.

(4) #s >>> #t

This relation holds if #s is instantiated by a node in the source c- or f-structure, #t is instantiated by a node in the target c- or f-structure, and those nodes are aligned.

Thus, the query in Example 5 will match all aligned pairs of analyses in a Norwegian-Dutch parallel treebank where a source c-structure lexical node *jente* is aligned with a target c-structure lexical node *meisje*.

(5) #s:"jente" >>> #t:"meisje"

An alignment relation can of course be part of a more complex query expression, as Example 6 illustrates. This query will find examples, like Example 2, where an argument is aligned with an adjunct, that is, aligned f-structures f1 (instantiating #f1) and f2, where a subsidiary argument f-structure s1 in f1 is aligned with a subsidiary adjunct f-structure s2 in f2: [5]

(6) #f1 >ARG #s1 & #f2 >(ADJ $) #s2
& #f1 >>> #f2 & #s1 >>> #s2

In a parallel treebank a single sentence in one language may correspond to multiple sentences in another. This can be handled on the overview Web page where manual alignment is implemented using drag and drop.

5 Conclusion

In this paper we have introduced the theoretical and methodological starting points for a linguistically motivated parallel treebanking approach that includes formal criteria for alignment. Rooting phrase alignment in correspondences at the level of predicate-argument structure within a parsebanking method which is both empirically founded and formally constrained offers a new approach to the study of the syntax-semantics interface across languages.

In the long run, this might open a new route to discovering language universals in this area, but currently we are only starting to explore this approach on a small number of typologically diverse languages. We have reported on the construction of a tool, a first prototype of which is operative and is being tested on test suites. Our work on alignment needs refinement and testing. We are also extending and testing the grammars in an integrated parsebanking approach and intend to move towards parsebanking of naturally occurring texts.

[5] '>ARG' is an abbreviation for '>(SUBJ | OBJ | ... | PREDLINK)', that is, the set of governable grammatical functions, '$' is the set-membership operator, and the expression '#f2 >(ADJ $) #s2' matches all pairs of f-structures f2, s2 where s2 is a member of the set constituting the value of ADJUNCT of f2.

References

[1] Joan Bresnan. *Lexical-Functional Syntax*. Blackwell, Malden, MA, 2001.

[2] Joan Bresnan and Lioba Moshi. Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry*, 21(2):147–185, 1990.

[3] Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. The Parallel Grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan*, 2002.

[4] Mary Dalrymple. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, San Diego, CA, 2001.

[5] Helge Dyvik. Translations as a semantic knowledge source. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pages 27–38, Tallinn, 2005. Institute of Cybernetics at Tallinn University of Technology, Institute of the Estonian Language.

[6] Anna Kibort. Extending the applicability of Lexical Mapping Theory. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG '07 Conference*, pages 250–270. CSLI Publications, Stanford, 2007.

[7] Paul Meurer. A computational grammar for Georgian. In *Proceedings of the Seventh International Tbilisi Symposium on Language, Logic and Computation*, Tbilisi, Georgia, 2007.

[8] Nazareth Amlesom Kifle. Differential object marking and topicality in Tigrinya. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG '07 Conference*, pages 5–25. CSLI Publications, Stanford, 2007.

[9] Victoria Rosén, Koenraad De Smedt, Helge Dyvik, and Paul Meurer. TREPIL: Developing methods and tools for multilevel treebank construction. In Montserrat Civit, Sandra Kübler, and Ma. Antònia Martí, editors, *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 161–172, 2005.

[10] Victoria Rosén, Paul Meurer, and Koenraad De Smedt. LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Frank Van Eynde, Anette Frank, Gertjan van Noord, and Koenraad De Smedt, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht, 2009. LOT.

[11] Yvonne Samuelsson and Martin Volk. Automatic phrase alignment: Using statistical n-gram alignment for syntactic phrase alignment. In Koenraad De Smedt, Jan Hajič, and Sandra Kübler, editors, *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories*, pages 139–150, 2007.